Next Token Prediction

Nipun Batra

July 21, 2025

IIT Gandhinagar

NEXT TOKEN GENERATION

— Inspired by <u>great</u> lecture(s)

from Andrej Karpathy.
↳ search for

Neural Networks
Zero to Hero

— We discussed relevance to chat GPT

app _?

what is the next character?

app _?

What is the next character?

Pose as classification task

app

c	P(c)
a	0.01
b	...
⋮	
e	0.4
⋮	
z	
_	

Specific Problem

- Generate Indian names

- Dataset : aabid
             aabida
             aadesh
             ⋮
             '
             '
             zeel

Specific Problem

- Generate Indian names

- Dataset : aabid
aabida
aadesh
$\vdots$
'
'
zeel

Assume

1) Only 26 lower
case char

2) _ indicates
end char

3) $4 < len < 10$

Generate Training Dataset

WORD #1

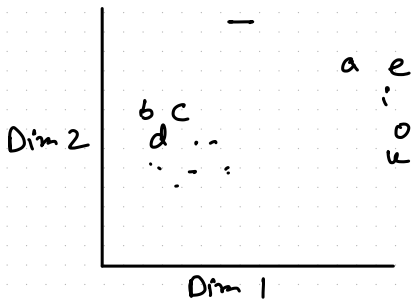 aabid

X			Y
say	we	consider	
−	−	−	a
−	−	a	a
−	a	a	b
a	a	b	i
a	b	i	d
b	i	d	−

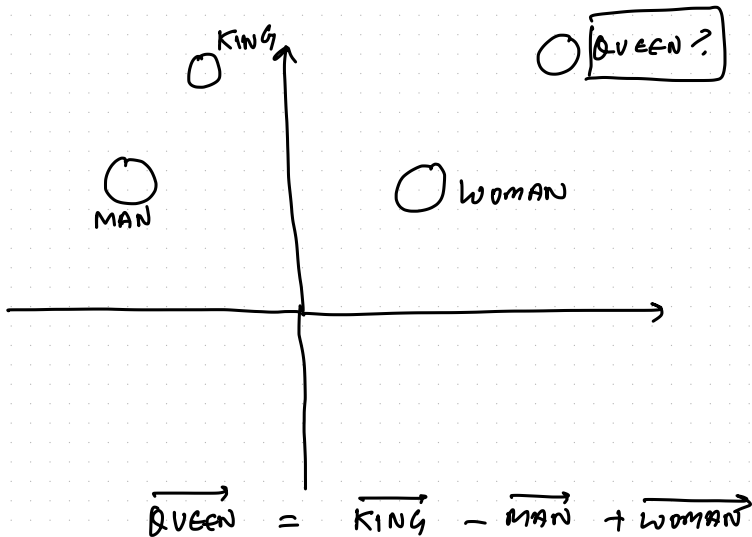'history / content' of 3 chars.

7 training examples from 1 name

Important Idea
Representation

→ Learn a vector representation for each character

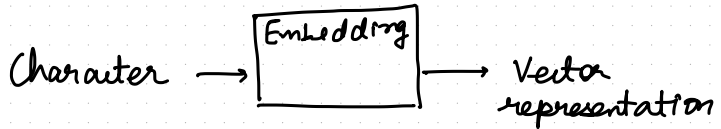→ 'Similar' characters → closer in vector space

WORD2VEC

KING

○ QUEEN?

○ WOMAN

MAN

$$\overrightarrow{QUEEN} = \overrightarrow{KING} - \overrightarrow{MAN} + \overrightarrow{WOMAN}$$

CHILD CRYING = CHILD SMILING + ADULT CRYING − ADULT SMILING

Embedding matrix | table

Character $\longrightarrow$ Embedding $\longrightarrow$ Vector representation

Embedding matrix | table

Character $\longrightarrow$ Embedding $\longrightarrow$ Vector representation
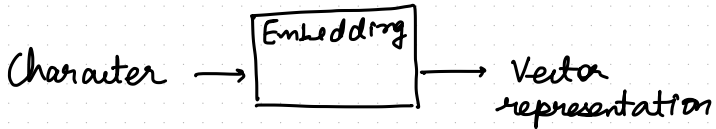
— Given 27 char $(a, .. z, \_)$, 'k' dim embedding

	Dim 1	Dim 2 ..	Dim K
a	0.1	.. .	..
b			
.			
.			
i			
z			
_			

Embedding matrix / table

Character $\longrightarrow$ [Embedding] $\longrightarrow$ Vector representation
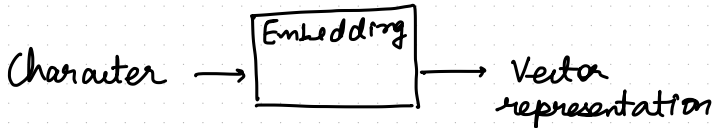
— Given 27 char $(a, .. z, -)$, 'k' dim embedding

	Dim1	Dim2 ..	Dim K
a	0.1	.. .	..
b			
:			
.			
i			
z			
—			

← LEARNABLE

OVERALL ARCHITECTURE

- For illustrat^2, 2dim embedding

X

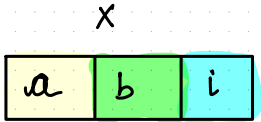| a	b	i

OVERALL   ARCHITECTURE

i) LOOKUP EMBEDDING

X

a	b	i

	D_1	D_2
a	0.1	0.3
b	-0.1	0.1
$\vdots$		
i	0.6	0.4
z		
$-$		

OVERALL ARCHITECTURE

i) LOOKUP EMBEDDING

X

a	b	i

	D_1	D_2
a	0.1	0.3
b	-0.1	0.1
...		
i	0.6	0.4
z		
_		

2) CONCATENATE EMBEDDINGS

X

a	b	i

	D_1	D_2
a	0.1	0.3
b	-0.1	0.1
⋮		
i	0.6	0.4
z		
—		

Feature Vector

0.1	0.3	-0.1	0.1	0.6	0.4

OVERALL ARCHITECTURE

3) MLP
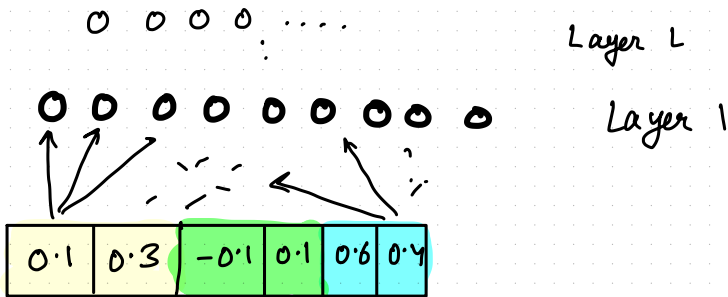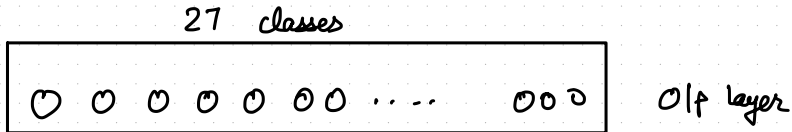
27 classes

Olp Layer

Layer L

Layer 1
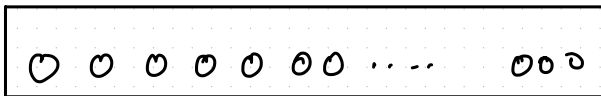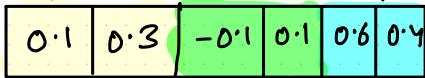
OVERALL ARCHITECTURE

4) USE CROSS ENTROPY LOSS TO LEARN
1) Embeddings
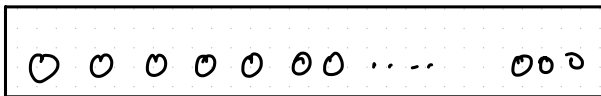2) MLP weights

27 classes

O/P Layer

Layer L

Layer 1

| 0.1 | 0.3 | -0.1 | 0.1 | 0.6 | 0.4 |

OVERALL ARCHITECTURE

4) USE CROSS ENTROPY LOSS TO LEARN
1) Embeddings
2) MLP weights

27 classes

$\bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \cdots \bigcirc \bigcirc \bigcirc$ O/P Layer

$\bigcirc \bigcirc \bigcirc \bigcirc \cdots$ Layer L

$\bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc$ Layer 1

| 0.1 | 0.3 | -0.1 | 0.1 | 0.6 | 0.4 |

5) GENERATION | SAMPLING

Test i|p

a b i

Test o|p

c	$p(c)$
a	0.01
b	0.01
c	0.01
d	0.6
e	.
.	.
.	.
.	.
.	.
z	(
—	.

5) GENERATION | SAMPLING

Test i/p

a b i

Sample from Prob.
distribution

a b i a   1%
a b i b   (1)%
a b i d   60%
⋮

Test o/p

c	p(c)
a	0.01
b	0.01
c	0.01
d	0.6
e	.
⋮	'
⋮	'
⋮	'
z	(
—	:

5) GENERATION / SAMPLING

aab → 1'/. → aaba → say 3'/. → aaba

aab → 60'/. → aabi